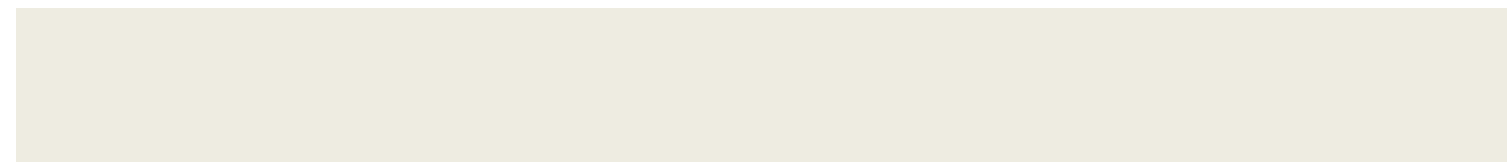


Fact-Checking



Presented By: MBBK

SOMMAIRE

1

Context

2

Présentation des données

3

Equilibrage des données

4

Features retenues

5

Prétraitement

6

Test et choix des meilleurs classifieurs

7

Matrice de confusion



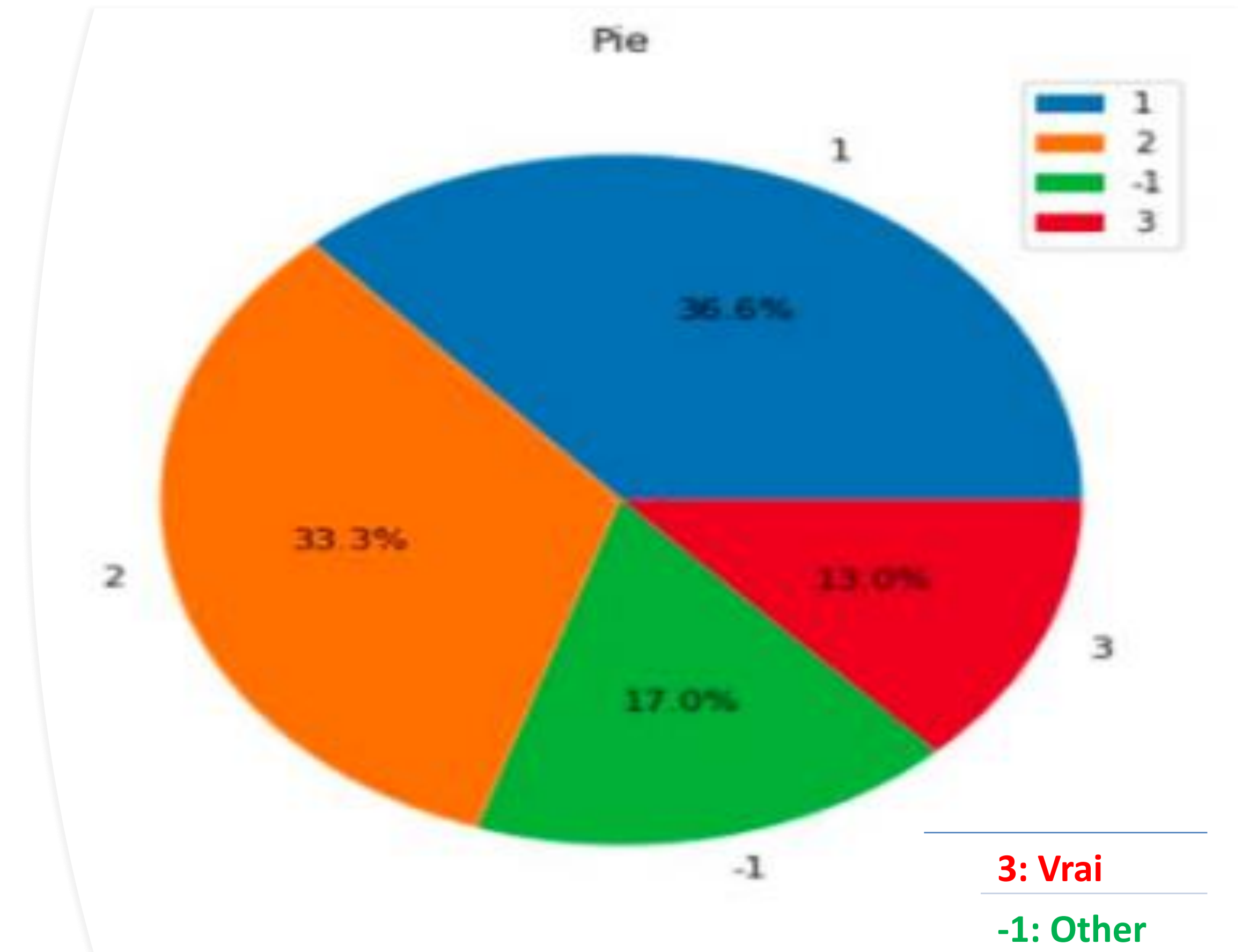
Context

- Ce projet de science des données consiste à faire du fact-checking de manière automatisé ou encore la classification d'assertions selon leur valeurs de véracité.
- Nous allons nous focaliser sur deux tâches de classification à savoir :
 - 1. Vrai vs Faux.
 - 2. Vrai et Faux vs Mixture.
- Dans les deux cas, la classification est binaire.

Présentation des données

Etude des
Différentes Données
DATASET

Utilisations de
différents
Graphiques



3: Vrai

-1: Other

1: Faux

2:Other

Equilibrage des données



OverSampling

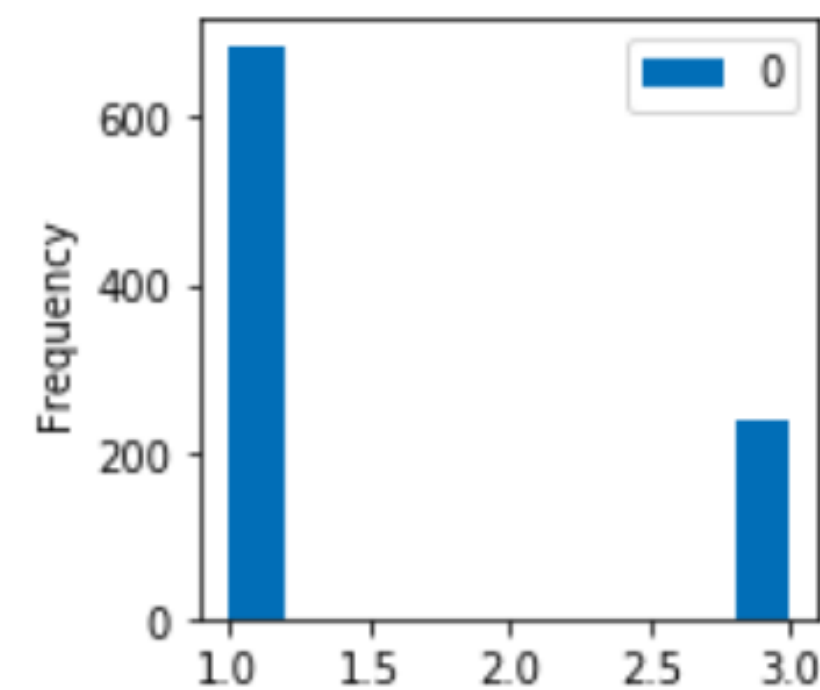
permet d'ajouter des données aux classes minoritaires pour le dataset {True vs False} la valeur de truthRating(3).



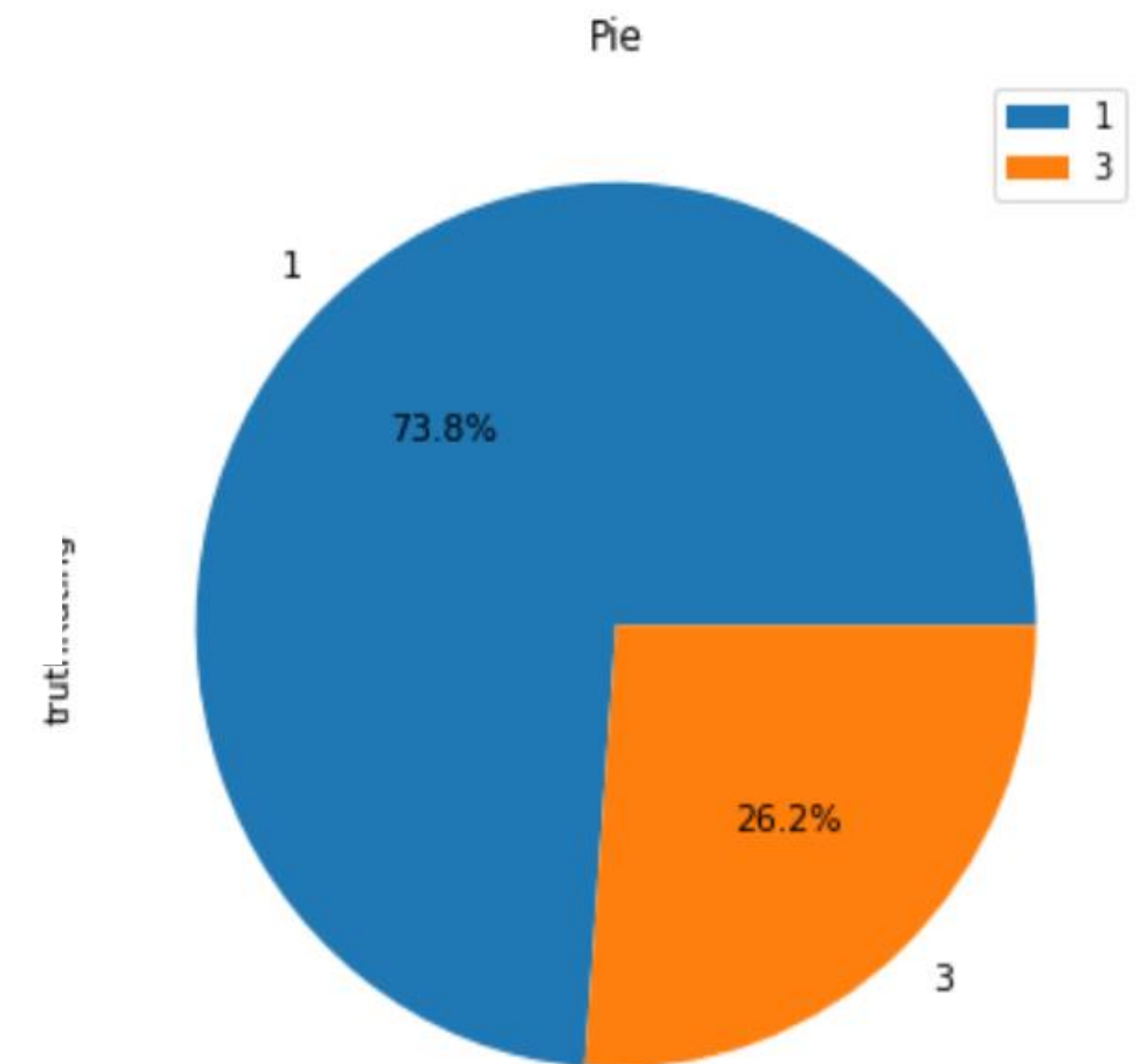
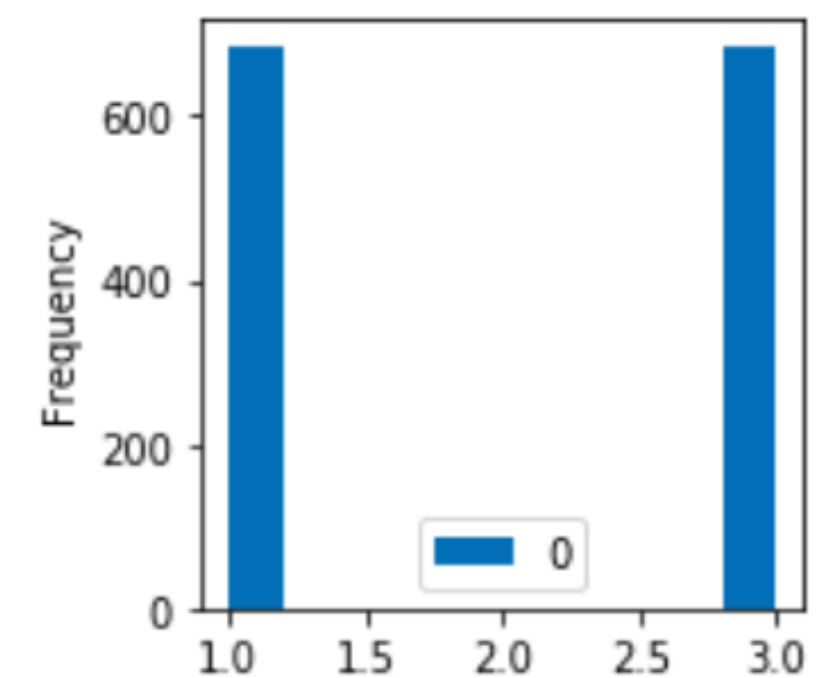
UnderSampling

consiste à enlever des données de la classe majoritaires, des assertions ayant comme valeurs de truth Rating 1 et 3

Avant



**Après
Oversampling**



Features retenues

✓ Text

✓ Author

✓ Source

✓ Keyword

✓ Headline

✗ ID

✗ Link

✗ Langage

✗ sourceURL

✗ Date

✓ Rating Name

✓ Author Headline

✓ Named entity claim

✓ Named entity article

Suppression Préliminaire

✓ Text

✓ Author

✓ Source

✓ Keyword

✓ Headline

✗ Rating Name

✗ Author Headline

✗ Named entity claim

✗ Named entity article

Suppression par test

✓ Text

✓ Author

✓ Source

✓ Keyword

✓ Headline

Les Features retenues

Prétraitement

Suppression des données manquantes

On cherche toutes les informations trouvées par les journalistes
ce cas nous intéresse
la rem
« unknown »

text author headline keywords source truthfulness

On cherche à enlever toutes les lignes où on trouve les valeurs vides keywords puisque dans ce cas la valeur vide ne nous intéresse pas pour la remplacer par « unknown » .

Prétraitement

```
=====ASCII pré-traitements=====
['Ted', 'Cruz', 'said', 'that', 'veterans', 'should', 'start', 'sell
ing', 'cookies', 'in', 'order', 'to', 'raise', 'funds', '.']
=====
=====
===== pré-traitements : mettre MINUSCULES pré-traitem
ents=====
['ted', 'cruz', 'said', 'that', 'veterans', 'should', 'start', 'sell
ing', 'cookies', 'in', 'order', 'to', 'raise', 'funds', '.']
=====
=====
===== pré-traitements : supprimer PONCTUATION pré-traiteme
nts =====
['ted', 'cruz', 'said', 'that', 'veterans', 'should', 'start', 'sell
ing', 'cookies', 'in', 'order', 'to', 'raise', 'funds']
=====
=====
```

Transformations sur les features

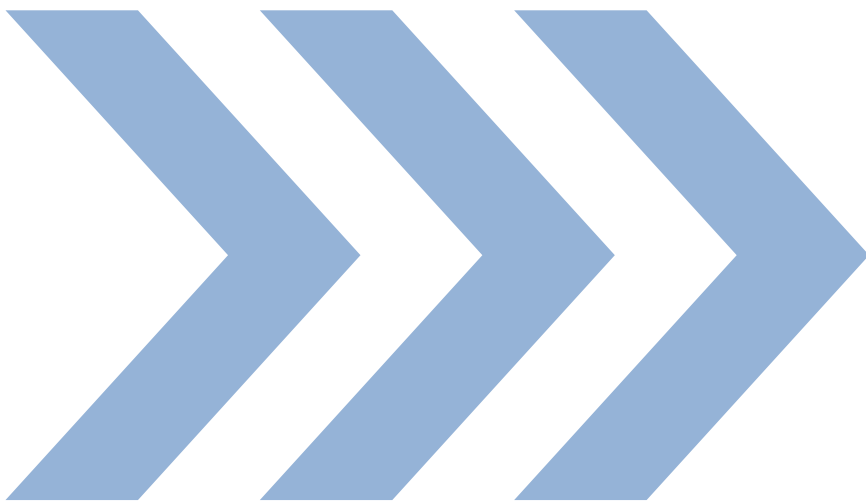
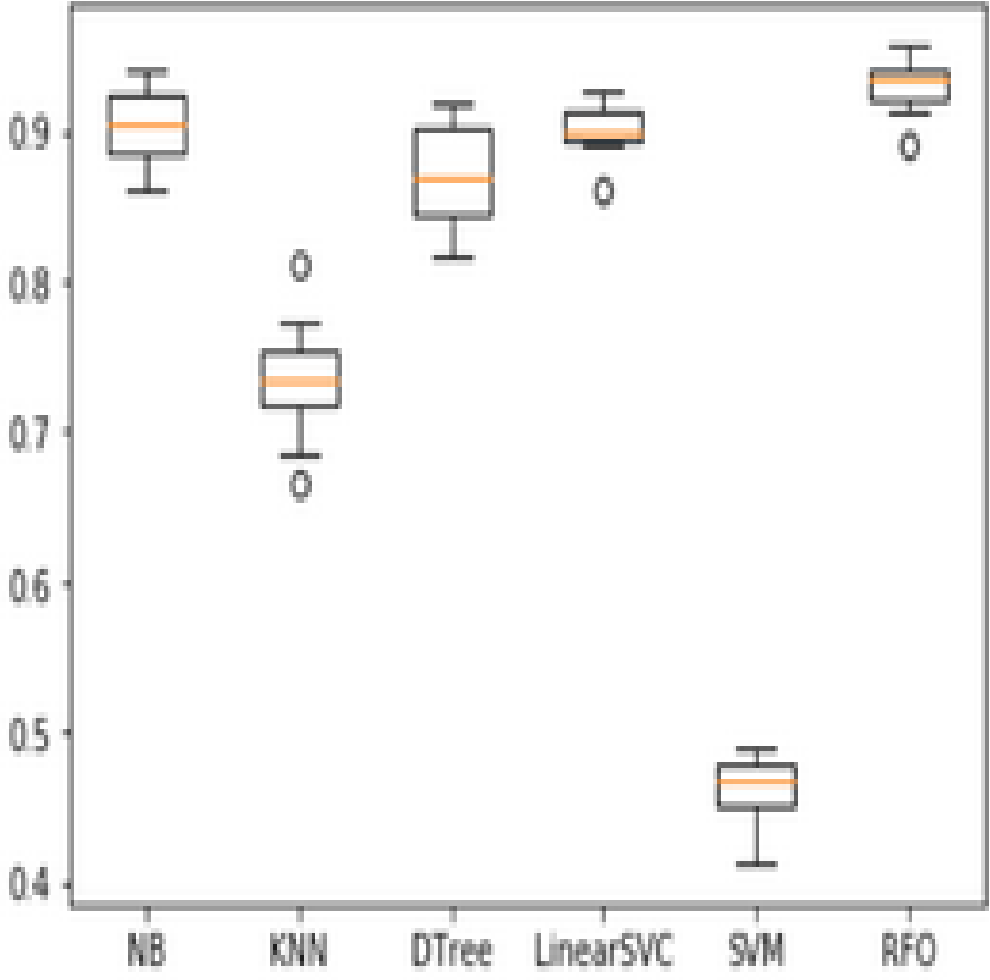
Liste des prétraitement

- ☒
- ☐ Suppression de caractères non-ASCII
- ☐ Convertir en Minuscule
- ☐ Supprimer les ponctuation et les caractères spéciale
- ☐ Remplacer les nombres
- ☐ Supprimer les espaces répéter
- ☐ Supprimer le stopwords
- ☐ Lemmatisation du texte
- ☐ Correction Orthographique
- ☐ Prendre en compte plusieurs mots consécutifs

Test et choix des meilleurs classifieurs

Tache de classification {True vs False}

Modèle	Score moyen	Déviations standard
RandomForest	93.3%	2.22%
GaussianNB	90,46%	2.53%
LinearSVC	90,17%	1.85%
DecisionTree	87.48%	3.63%
KNN	73,5%	3.9%
SVM	46.29%	2.13%



Meilleur classifieurs

RandomForest

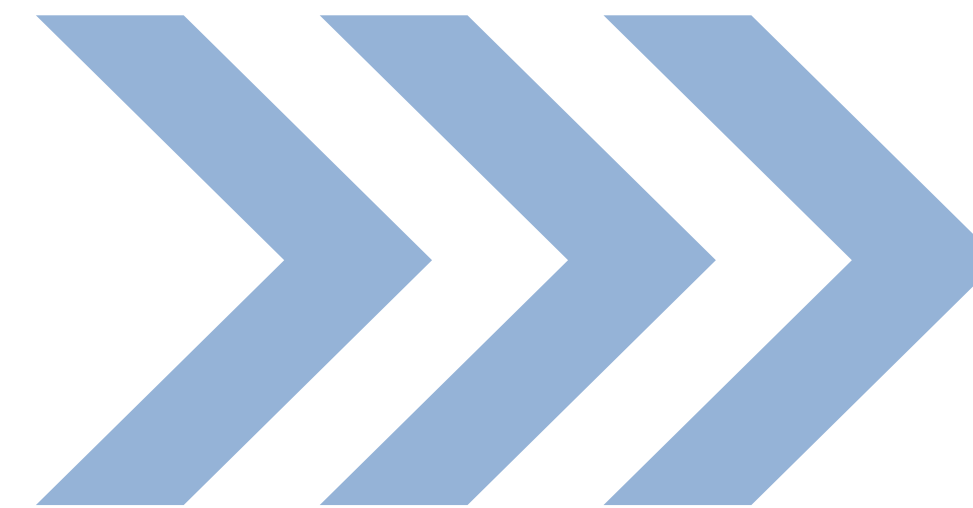
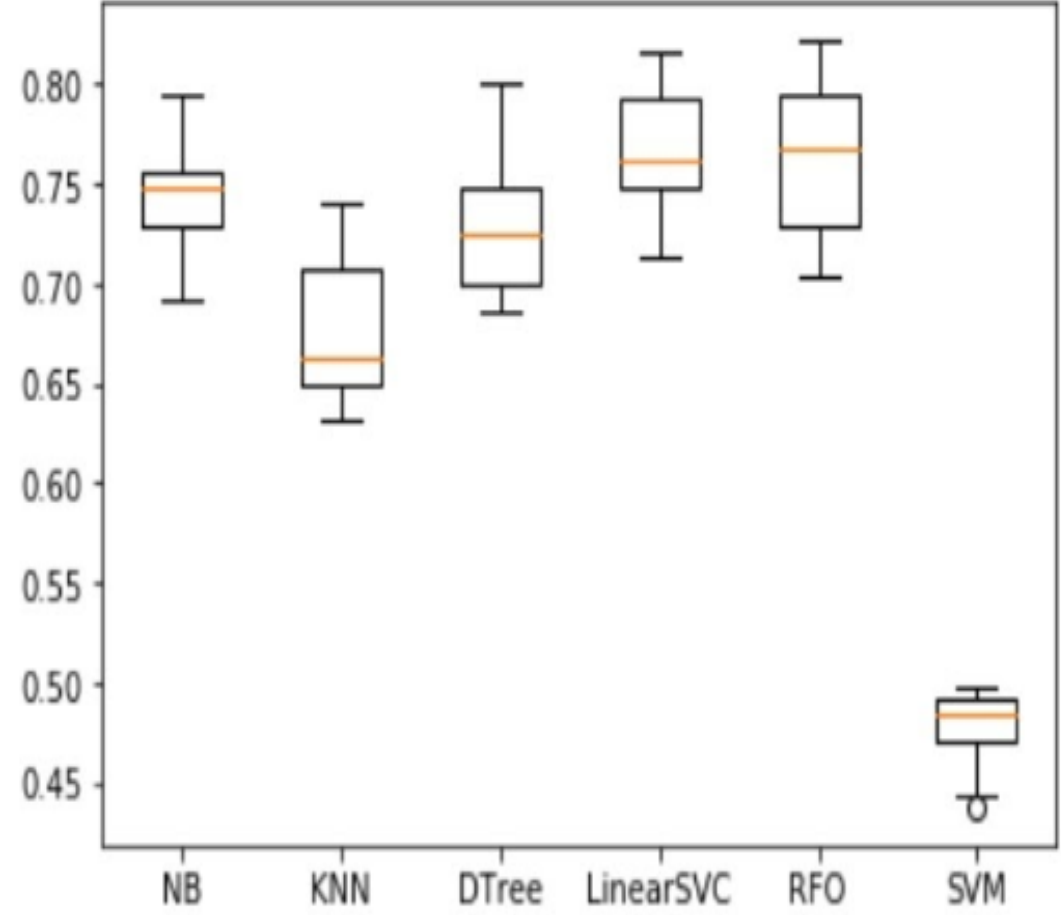
Gaussian Naive Bayes

Test et choix des meilleurs classifieurs

10

Tache de classification {True et False} vs {Mixture}

Modèle	Accuracy moyen	Déviation standard
Random Forest	76.43%	38.61%
LinearSVC	76.86%	31.77%
GaussianNB	74.32%	2.96%
Decision Tree	72.81%	3.36%
KNN	67.83%	3.78%
SVM	47.72%	20.23%

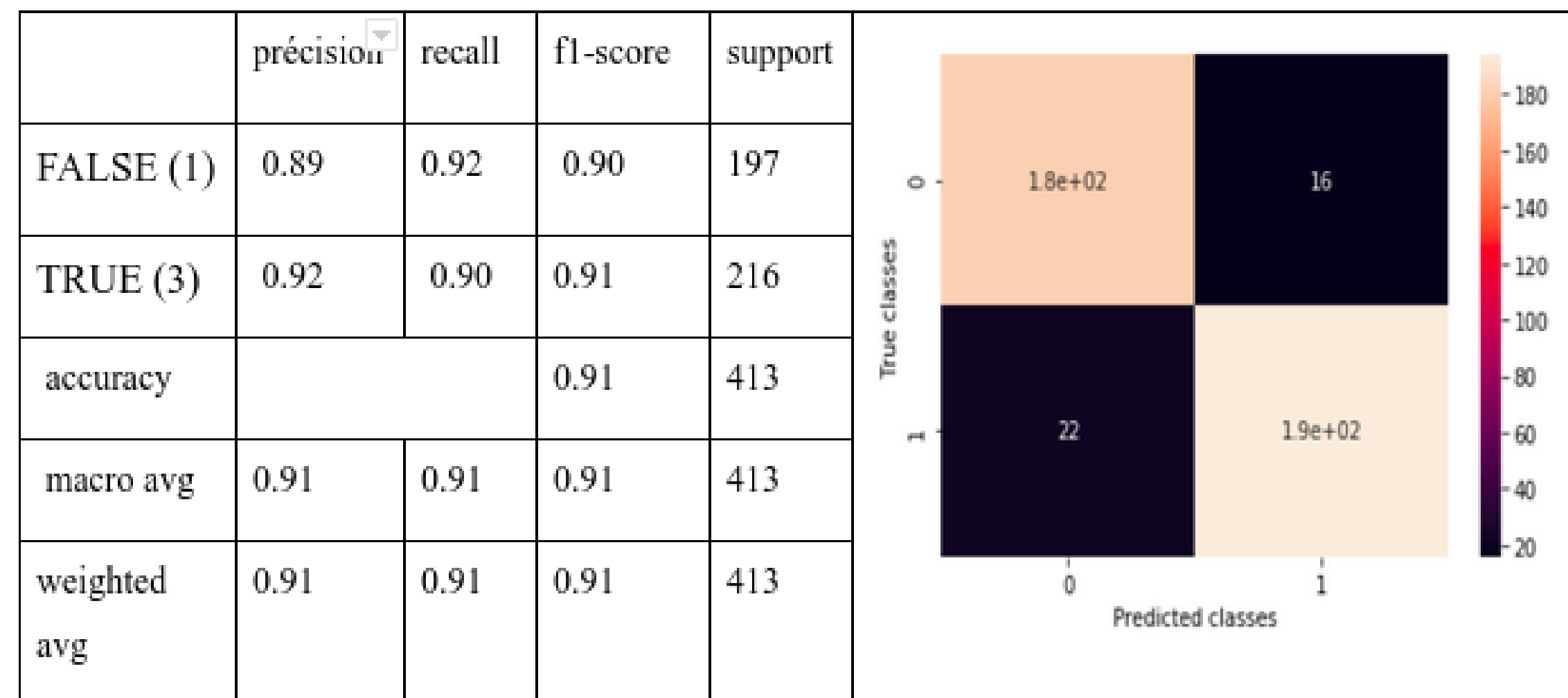


Meilleur classifieurs

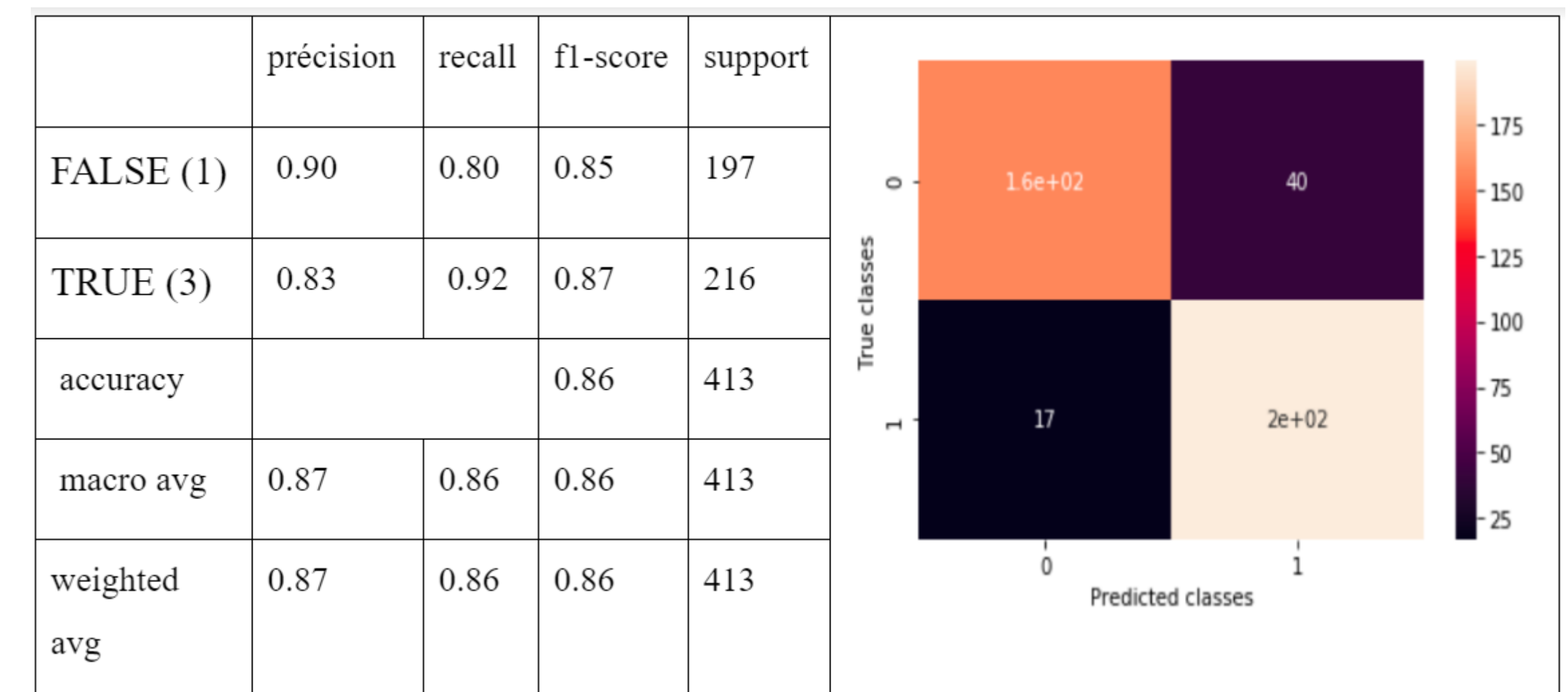
GaussianNB

Random Forest

Matrice de confusion



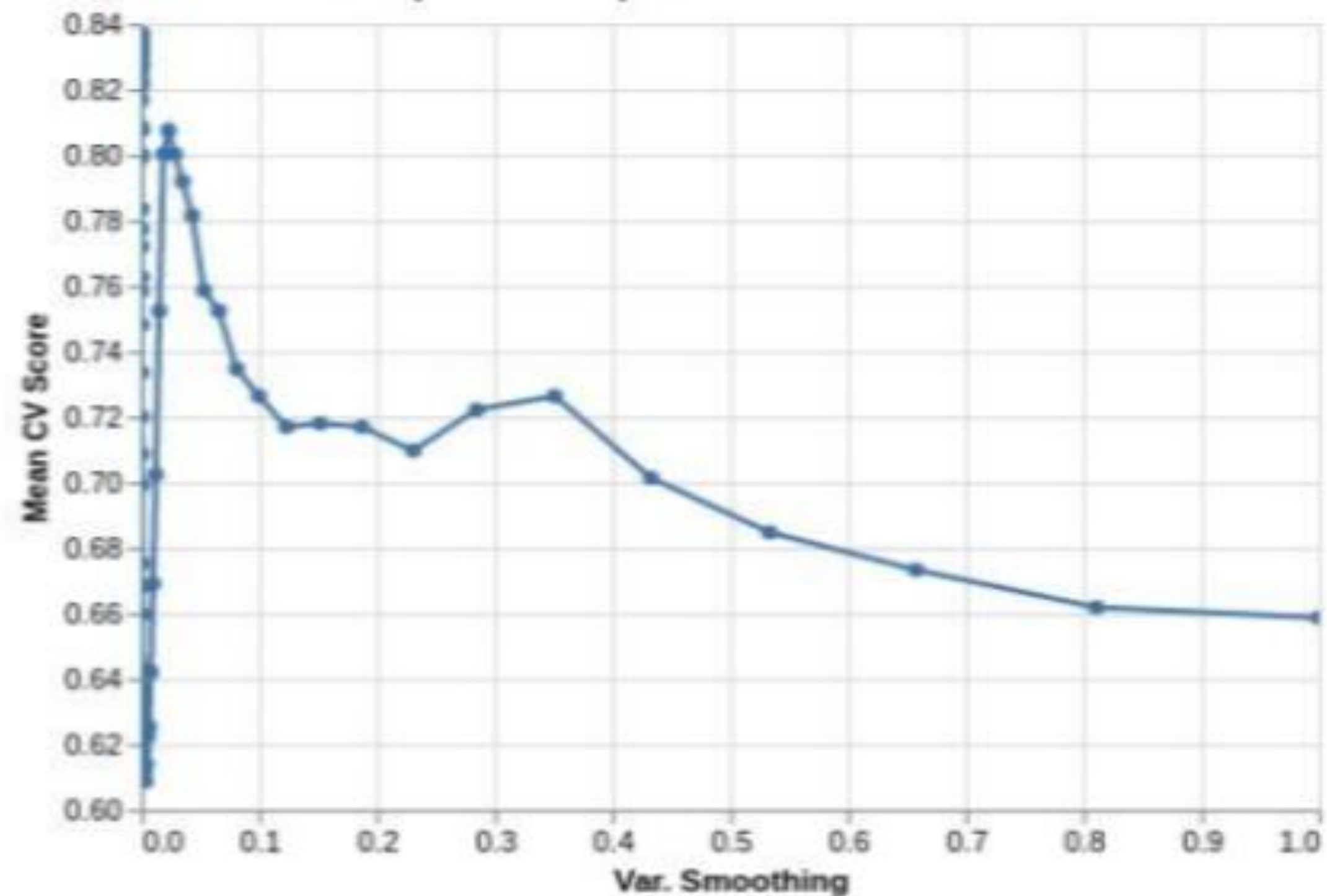
Naive Bayes {True} vs {False}



Random Forest {True} vs {False}

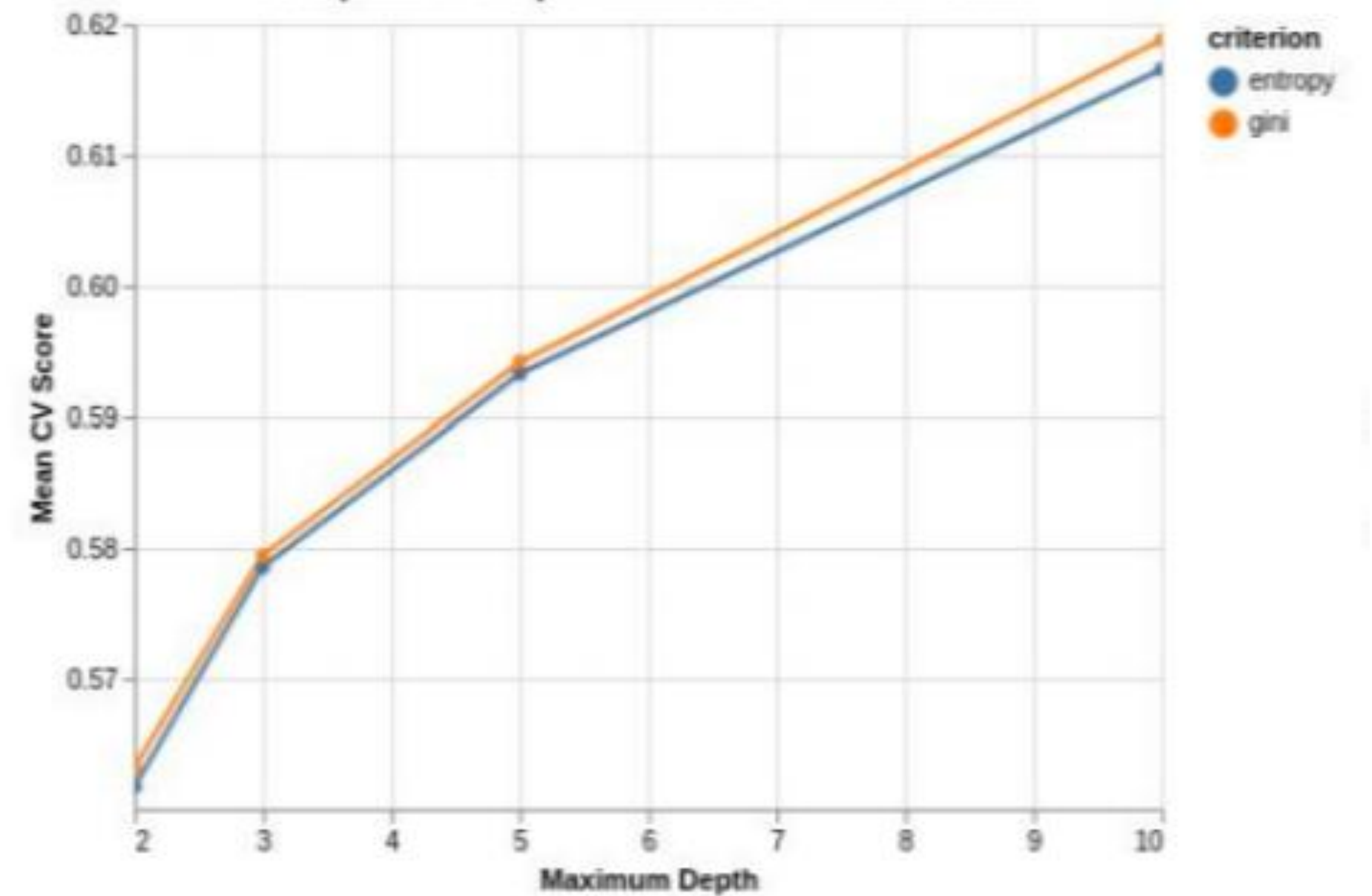
Grid Search

Compraison de performance de GaussinNB



Grphe GridSearch Naive Bayes {True} vs {False}

Comparaison de performance de RandomForest



Grphe GridSearch Random Forest {True} vs {False}

Conclusion

Après la sauvegarde on a essayer de prédire des vrais Données, alors les résultats de prédiction sont pas mal en comparant avec les résultats réelle.

réelles

assertion : 4	réelle	1	prédite	1
assertion : 7	réelle	1	prédite	1
assertion : 9	réelle	1	prédite	1
assertion : 10	réelle	3	prédite	1
assertion : 11	réelle	1	prédite	1
assertion : 19	réelle	1	prédite	1
assertion : 21	réelle	1	prédite	1
assertion : 22	réelle	1	prédite	1
assertion : 24	réelle	1	prédite	1
assertion : 25	réelle	1	prédite	1
assertion : 26	réelle	1	prédite	1
assertion : 27	réelle	1	prédite	1
assertion : 28	réelle	1	prédite	1
assertion : 30	réelle	1	prédite	1
assertion : 31	réelle	3	prédite	1



Merci de votre intention